Vergleich unterschiedlicher Ansätze zur instrumentellen Vorhersage der Qualität von Text-to-Speech Systemen: Daten der Blizzard Challenge 2010

Florian Hinterleitner[1], Sebastian Möller[1],
Christoph Norrenbrock[2], Ulrich Heute[2]

[1] *Quality and Usability Lab, Deutsche Telekom Laboratories, TU Berlin, Deutschland,*

[2] *Digital Signal Processing and System Theory, CAU Kiel, Deutschland*

florian.hinterleitner@telekom.de, sebastian.moeller@telekom.de,
cno@tf.uni-kiel.de, uh@tf.uni-kiel.de

Einleitung

In den letzten Jahren hat die Entwicklung von Sprachsynthesen große Fortschritte gemacht. Sogenannte Text-to-Speech-(TTS)-Systeme klingen inzwischen natürlich genug, um sie in vielen gängigen Anwendungen einzusetzen. Die Generierung natürlich klingender Sprechrhythmen, Intonationen und Akzente bleibt jedoch nach wie vor die für die Entwickler von Sprachsynthesen größte Herausforderung. Um den Aufwand für die sowohl kosten- als auch zeitintensiven Hörversuche, die zur Bestimmung der Qualität von Synthesen nötig sind, auf ein Minimum zu reduzieren, ist ein instrumentelles Verfahren zur Vorhersage der Qualität wünschenswert.
Unterschiedliche Ansätze wurden 2010 auf deutsch- und englischsprachige Datenbanken trainiert und erzielten für diese teilweise erstaunlich genaue Qualitätsvorhersagen [1] [2]. Im Folgenden werden die Ergebnisse der Ansätze auf dem Datensatz der Blizzard Challenge (BC) 2010 untersucht.

Vorhersagemodelle

In diesem Abschnitt werden die zur Qualitätsvorhersage verwendeten Schätzer vorgestellt. Dabei handelt es sich zum einen um einen HMM-basierten Ansatz [3], zum anderen um ein Modell, das die internen Parameter des Schätzers nach ITU-T Empfehlung P.563 [4] verwendet. Beide Modelle sind in Abbildung 1 zu sehen.

HMM-basierter Ansatz

Wie im rechten Teil der Abbildung 1 zu sehen ist, besteht der Algorithmus aus zwei nacheinander ausgeführten Prozessen. Zunächst wird ein Hidden-Markov-Modell (HMM) auf natürliche Sprache trainiert (gestrichelte Linien), dieses dient in der anschließenden Bewertungs-Phase (durchgezogene Linien) als Vergleichsmaß für die TTS-Signale. Sowohl das natürliche als auch das synthetische Signal durchlaufen bis dorthin die gleiche Vorverarbeitung und Merkmalsextraktion.

Vorverarbeitung und Merkmalsextraktion

Zunächst findet bei beiden Signalen eine Pegelanpassung mittels eines Active Speech-Level Meters auf einen mittleren Pegel bei Sprachaktivität von -26dB relativ zur Aussteuerungsgrenze statt. Da zur Bewertung der Qualität nur Abschnitte des Signals mit aktiver Sprache von

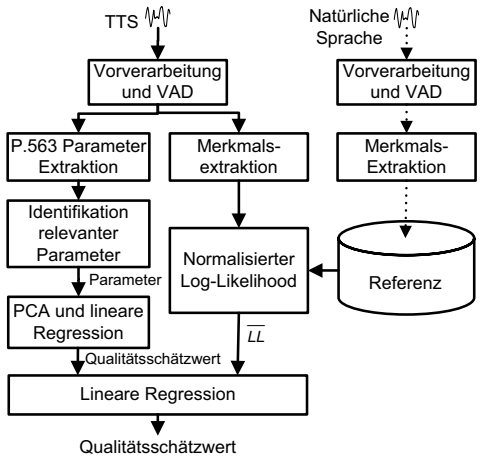

Abbildung 1: Modelle zur Qualitätsvorhersage von TTS-Systemen

Nutzen sind, werden Stillephasen und Pausen, mit einer Länge von über 75ms, durch einen simplen Energieschwellwert Voice-Activity-Detection-Algorithmus aus dem Signal entfernt. Während der Merkmalsextraktion werden, bei einer Fenstergröße von 25ms und einer zeitlichen Verschiebung von 10ms, MFCCs 12. Ordnung berechnet. Der MFCC-Koeffizient Nullter Ordnung wird als logarithmisches Energiemaß verwendet.

HMM-Training

Die aus den natürlichen Sprachsignalen extrahierten Merkmale werden zum Training der HMMs verwendet. Das eingesetzte HMM-Modell besteht aus 8 Zuständen, wobei die Verteilungen der Ausgangswerte jedes Zustands aus gemischten Gaußverteilungen mit 16 gaußschen Diagonal-Kovarianz-Komponenten bestehen. Übergangswahrscheinlichkeiten, initiale Wahrscheinlichkeiten, Beobachtungsverteilungsdichten und sonstige Modellparameter werden mittels des Expectation-Maximization-Alogrithmus berechnet.

Berechnung des normalisierten Log-Likelihood (LL)

Für die aus den TTS-Signalen extrahierten Merkmale können mittels der trainierten HMMs und dem Forward-Backward-Algorithmus LL-Werte gewonnen werden. Diese werden auf die Anzahl der im Signal vorkommenden

	Blizzard Challenge											
	2008 (Full Corpus)			2009 (Full Corpus)			2010 (Full Corpus)			2010 (Arctic Corpus)		
	R	ϵ	ρ	R	ϵ	ρ	R	ϵ	ρ	R	ϵ	ρ
LL	0,64	21,62	0,53	0,17	21,15	0,14	0,42	21,15	0,37	-0,53	21,78	-0,32
P.563	0,64	0,45	0,67	**0,65**	0,52	**0,69**	0,29	0,65	0,02	**0,61**	0,53	**0,44**
LL + P.563	**0,79**	0,41	**0,75**	0,59	0,52	0,50	**0,50**	0,59	**0,44**	-0,22	0,64	-0,10

Tabelle 1: Korrelationen zwischen auditiver Bewertung und prognostizierter Qualität

aktiven Sprachabschnitte normalisert. Der normalisierte LL gilt als Maß für die Qualität des TTS-Signals und wird im Folgenden mit $\overline{LL}$ bezeichnet.

Interne Parameter des P.563

Als Basis des zweiten Ansatzes dienen die internen Störungsparameter des P.563. Dieses Modell wird zur referenzlosen Qualitätsvorhersage telefonkanalkodierter Sprache verwendet. Zur Ermittlung des Qualitätsschätzwerts berechnet P.563 einige interne Parameter, die unter anderem Störungen wie Rauschen, Clipping oder Robotereffekte wiedergeben. Aus den 44 berechneten Parametern wurden diejenigen ausgewählt, die für die vorliegenden Datenbanken im Mittel eine Korrelation mit den auditiven Bewertungen von $|\rho| \geq 0,40$ aufwiesen. Eine anschließende Hauptkomponentenanalyse lieferte vier Faktoren. Diese wurden mittels linearer Regression, mit dem auditiven Natürlichkeitsurteil als Zielvariable, zu einem Qualitätsschätzer kombiniert.

Schließlich lies sich durch eine weitere lineare Regression, aus dem $\overline{LL}$-Ansatz und dem P.563-Modell ein Schätzer erzeugen, der beide Ansätze miteinander kombiniert.

Datenbanken der Blizzard Challenges 2008, 2009 und 2010

Die vorgestellten Verfahren wurden auf Daten der BCs 2008 und 2009 trainiert und auf den unabhängigen Daten von 2010 getestet. Teilnehmer dieser Wettbewerbe erhalten einen 15 stündigen englischen Sprachkorpus (Full Corpus) auf den sie ihre Synthesesysteme trainieren können. Weitere Aufgabenstellungen geben kleinere Korpora vor: den Arctic Corpus mit 1h Sprachdaten und der Small Corpus bestehend aus 10 bis 100 Sätzen des Arctic Corpus.
Für das Training der HMMs wurden die natürlichen Referenzstimuli aus dem jeweiligen Datensatz verwendet.

Ergebnisse

Die auditiven Bewertungen der einzelnen Stimuli wurden für jedes der Systeme gemittelt. Die Korrelationen zwischen diesen Werten und den von den vorliegenden Modellen für die einzelnen Systeme prognostizierten Werten sind in Tabelle 1 zu sehen. Neben dem Korrelationskoeffizienten nach Pearson R und Spearmans Rangkorrelationskoeffizient ρ ist ebenfalls die Wurzel des mittleren quadratischen Fehlers ϵ angegeben.
Für das HMM-Modell ergeben sich für die Daten von 2008 sowie für die Daten des Full Corpus von 2010 deutlich positive Korrelationen. Auf dem 2009er Datensatz liegen sowohl R als auch ρ nur noch leicht im positiven Bereich. Auf den Daten des Arctic Corpus von 2010 versagt dieser Ansatz.
Das Modell der P.563 Störungsparameter zeigt auf den Trainingsdaten (BC 2008 und 2009) Korrelationen von über 0,64. Auf dem unabhängigen Datensatz von 2010 liegt es für den Full Corpus noch bei 0,28, für die Daten des Arctic Corpus wird sogar ein R von 0,62 erreicht.
Durch eine Kombination der beiden Ansätze lässt sich für BC 2008 sowie für BC 2010 (Full Corpus) noch einmal eine deutliche Steigerung der Vorhersagequalität feststellen.

Fazit

Die Ergebnisse zeigen, dass eine Qualitätsvorhersage für TTS-Daten nur auf Basis des synthetisierten Sprachsignals möglich ist. Für die beiden Ansätze ergeben sich jedoch je nach Datensatz deutliche Schwankungen in der Genauigkeit der Vorhersage. Eine Analyse dieser Mängel im Bezug auf den Einfluss der Trainingsdaten für das HMM-Modell, die Auswahl der Merkmale für den P.563-Schätzer sowie den Einfluss der Synthesesysteme wird angestrebt.

Literatur

[1] Möller, S. ; Hinterleitner, F. ; Falk, T.H. ; Polzehl, T.: Comparison of Approaches for Instrumentally Predicting the Quality of Text-To-Speech Systems. In: *Proceedings of the 11th Annual Conference of the ISCA (Interspeech 2010). International Speech Communication Association (ISCA)* (2010)

[2] Hinterleitner, F. ; Möller, S. ; Falk, T.H. ; Polzehl, T.: Comparison of Approaches for Instrumentally Predicting the Quality of Text-to-Speech Systems: Data from Blizzard Challenges 2008 and 2009. In: *Proceedings of the Blizzard Challenge Workshop. International Speech Communication Association (ISCA)* (2010)

[3] Falk, T. H. ; Möller, S.: Towards Signal-Based Instrumental Quality Diagnosis for Text-to-Speech Systems. In: *IEEE Signal Processing Letters* 15 (2008), S. 781–784

[4] ITU-T Rec. P.563: *Single Ended Method for Objective Speech Quality Assessment in Narrow-Band Telephony.* Geneva: International Telecommunication Union, 2004